Applied Statistics
Programming Assignment

In this project, we will try to predict the selling price (in dollars) of houses in the Saratoga suburb in NY, based on the following variables:

- The size of the house (in square feet).
- The number of bathrooms.
- The number of bedrooms.
- Whether the house has a fireplace.
- The number of acres of the house plot.
- The age of the house.

The data are available in the course website. You can implement the assignment in any programming language you want. Your main deliverable is a report, written in Word or similar editor, that includes your answers to questions below, which can be supported by plots and/or results copied from your analysis. Your report should have the following components: An introductory paragraph, describing the goals of your analysis, and one section per exercise. Each section must summarize your findings and conclusions for the corresponding exercise.
You must also include your code, but I will not necessarily run it. Do not copy code, because I will run it through code similarity detection software. However, it is ok to collaborate with your colleagues and compare your results. You can submit your document and accompanying code here on the course website, by Monday, June 27th, 23:59. Late submissions will not be accepted.

Exercise 1: (20 points) Produce scatter plots and compute the correlation coefficient for every one of your independent (explanatory) variables against your dependent (response) variable (house price). If you could only use one variable to predict house prices, which one would you use?

Exercise 2: (60 points) Use multiple regression and use all of your explanatory variables to predict house prices.

(a) Are the conditions for the multiple linear regression model satisfied? Explain your answers.

(b) What is the slope associated with the variable you identified as most predictive in Exercise 1? What is the interpretation of the slope?

(c) Which variables are significant for the prediction of house prices?

Exercise 3: (20 points) Implement the backward elimination or the forward selection algorithm to select the most important variables for predicting house prices. Which variables does your final model include? Write your conclusions.

Εφαρμοσμένη Στατιστική
Προγραμματιστική Άσκηση

Σε αυτή την εργασία, θα προσπαθήσουμε να προβλέψουμε την τιμή πώλησης (σε δολάρια) των σπιτιών στο προάστιο Saratoga στη Νέα Υόρκη, με βάση τις ακόλουθες μεταβλητές:

- Το μέγεθος του σπιτιού (σε τετραγωνικά πόδια).
- Ο αριθμός των μπάνιων.
- Ο αριθμός των υπνοδωματίων.
- Αν το σπίτι έχει τζάκι.
- Ο αριθμός των στρεμμάτων του οικοπέδου του σπιτού.
- Η ηλικία του σπιτιού.

Θα βρείτε τα δεδομένα στο σάιτ του μαθήματος. Μπορείτε να υλοποιήσετε την εργασία σε οποιαδήποτε γλώσσα προγραμματισμού θέλετε. Το κύριο παραδοτέο σας είναι μια αναφορά, γραμμένη σε Word ή παρόμοιο πρόγραμμα επεξεργασίας, που περιλαμβάνει τις απαντήσεις σας στις παρακάτω ερωτήσεις, οι οποίες θα υποστηρίζονται από γραφικές παραστάσεις ή/και άλλα αποτελέσματα του κώδικά σας. Η αναφορά σας θα πρέπει να έχει τα ακόλουθα στοιχεία: Μια εισαγωγική παράγραφο, που περιγράφει τους στόχους της ανάλυσής σας, και μια ενότητα ανά άσκηση. Κάθε ενότητα πρέπει να συνοψίζει τα ευρήματά σας και τα συμπεράσματά σας για την αντίστοιχη άσκηση.
Πρέπει επίσης να συμπεριλάβετε τον κώδικά σας, αλλά δεν θα τον εκτελέσω απαραίτητα. Μην αντιγράψετε κώδικα, γιατί θα χρησιμοποιήσω πρόγραμμα ανίχνευσης ομοιότητας λογισμικού. Ωστόσο, είναι εντάξει να συνεργάζεστε με τους συναδέλφους σας και να συγκρίνετε τα αποτελέσματά σας. Μπορείτε να υποβάλετε το αρχείο σας και τον συνοδευτικό κωδικα στο σαιτ του μαθήματος, έως τη Δευτέρα, 27 Ιουνίου, 23:59. Εκπρόθεσμες υποβολές δεν θα γίνονται δεκτές.

Άσκηση 1: (20 βαθμοί) Δημιουργήστε διαγράμματα διασποράς και υπολογίστε τον συντελεστή συσχέτισης για κάθε μία από τις ανεξάρτητες (επεξηγηματικές) μεταβλητές σας έναντι της εξαρτημένης μεταβλητής (απόκρισης) (τιμή κατοικίας). Αν μπορούσατε να χρησιμοποιήσετε μόνο μία μεταβλητή για να προβλέψετε τις τιμές των κατοικιών, ποια θα χρησιμοποιούσατε;

Άσκηση 2: (60 βαθμοί) Χρησιμοποιήστε πολλαπλή γραμμική παλινδρόμηση και χρησιμοποιήστε όλες τις επεξηγηματικές σας μεταβλητές για να προβλέψετε τις τιμές των κατοικιών.

(a) Ικανοποιούνται οι προϋποθέσεις για το μοντέλο πολλαπλής γραμμικής παλινδρόμησης; Δικαιολογήστε τις απαντήσεις σας.

(b) Ποια είναι η κλίση που σχετίζεται με τη μεταβλητή που προσδιορίσατε ως πιο σημαντική στην Άσκηση 1; Ποια είναι η ερμηνεία της κλίσης;

(c) Ποιες μεταβλητές είναι σημαντικές για την πρόβλεψη των τιμών των κατοικιών;

Άσκηση 3: (20 βαθμοί) Εφαρμόστε τον αλγόριθμο εξάλειψης προς τα πίσω (backward elimination) η τον αλγόριθμο έμπροσθεν επιλογής (forward selection) για να επιλέξετε την πιο σημαντική μεταβλητή για την πρόβλεψη των τιμών των κατοικιών. Ποιες μεταβλητές περιλαμβάνει το τελικό μοντέλο σας; Τι συμπεράσματα βγάζετε?